REMARKS

The drawings were objected to for several reasons. First, reference number 195 of Fig. 1 was not found in the specification. To correct this, Applicants have amended the specification on page 9 to correct an obvious error where output peripheral interface 195 was identified by reference number 190 instead of 195.

Second, Fig. 3 was objected to for using reference number 306 twice to refer to the same item. A corrected Figure 3 is submitted herewith, which corrects this duplication.

Third, Fig. 4 was objected to because item 402 was not labeled. Fig. 4 provides an example of a display screen. The text shown in Fig. 4, such as "Record Pronunciation", is text that would be displayed to a user. Box 402 in Fig. 4 is an input text box. As such, it remains empty until the user types in a new word. Thus, text box 402 should not be labeled in Fig. 4.

The specification was objected to because the reference to "190" on line 31 of page 9 should have been "195". As noted above, this has been corrected with the present amendment. In addition, claim 18 was objected to because it used the word "a" where "an" should have been used. With the present amendment, claim 18 has been amended to correct this error.

Claims 1-4

Claims 1-4 were rejected under 35 U.S.C. § 102(a) as being anticipated by Smith et al. (U.S. Patent 6,408,271 B, hereinafter Smith).

Smith discloses a system for generating possible pronunciations of a sequence of words. Under Smith, each word has many possible pronunciations. As a result, for a sequence of words, there are multiple possible combinations of these pronunciations. Smith selects the top N pronunciations for the sequence of words to store in a dictionary and use during speech recognition. During speech recognition, a decoder compares input

feature vectors to pronunciations in the dictionary to determine if any of the pronunciations match the user's speech.

Independent claim 1 provides a method of adding an acoustic description of a word to a speech recognition lexicon. Initially, the text of a word is converted into an orthographically derived acoustic description of the word. The orthographically derived acoustic description is then scored based in part on a comparison between the orthographically derived acoustic description and a speech signal representing a user's pronunciation of the word. The speech signal is also decoded to produce a decoded acoustic description of the word and a score for the decoded acoustic description. One of the orthographically derived acoustic description and the decoded acoustic description is then selected as the acoustic description of the word based on the scores for the two acoustic descriptions.

Smith does not show or suggest the invention of claim 1 because it does not show or suggest steps of decoding a speech signal representing a user's pronunciation of a word to produce a decoded acoustic description of the word.

In the Office Action, it was asserted that steps 202 and 302 in combination with speech signal 804 showed the step of decoding a speech signal to form an acoustic description and column 12, lines 26-37 showed the steps of selecting between that description and another description. Applicants respectfully dispute these assertions.

First, steps 202 and 302 do not decode a speech signal to identify an acoustic description. Instead, the acoustic description is formed though text-to-phoneme rules or by using a transcription dictionary. (See Smith, col. 6, lines 16-20). Smith does not teach that steps 202 or 302 use a speech signal in any manner. Further it is clear that Smith does not generate an acoustic description from the speech signal because without the

text-to-phoneme rules there would be no acoustic descriptions in Smith.

Second, column 12, lines 26-37 do not show a step of selecting between a decoded acoustic description that is formed from a speech signal and an orthographically derived acoustic description that is formed from text. In the cited section, Smith simply discusses generating possible pronunciations for a sequence of words based on possible pronunciations for the individual words in the sequence. All of these pronunciations are generated from text as stated in column 6, lines 16-20. Thus, none of the pronunciations is decoded from a speech signal.

Since Smith does not show the steps of generating a decoded acoustic description from a speech signal or selecting between such a decoded acoustic description and another acoustic description, it does not anticipate claim 1 or claims 2-4, which depend therefrom.

Claim 5

Claim 5 was rejected under 35 U.S.C. § 103(a) as being obvious from Smith in view of Bahl et al. (U.S. Patent 5,875,426, hereinafter Bahl '426).

Bahl '426 provides a speech recognition system that is able to handle word pronunciations that are context dependent. During recognition, Bahl '426 first considers all possible stored pronunciations for all words in a vocabulary. The speech signal is applied to these pronunciations to identify a set of candidate words. These candidate words are applied to a language model that generates a score for each current candidate word based on a previously identified word. This results in a ranked list of candidate current words and the dictionary-based pronunciations of those words.

Bahl '426 then examines a field in each current word's dictionary entry and a field in the preceding word's dictionary entry to determine if an additional pronunciation of the word

should be added as a candidate. Note that this additional pronunciation candidate is a rule-based candidate and is not dependent on how the speaker pronounced the word. The speech signal is then applied to these candidate words and pronunciations in order to select a most likely word.

Dependent claim 5 depends from claim 1 and includes a further limitation wherein decoding a speech signal to form a decoded acoustic description further comprises using a language model.

Because claim 5 depends from claim 1, it includes the limitation to producing a decoded acoustic description of a word by decoding a speech signal representing the user's pronunciation of the word. Neither Smith nor Bahl '426 show a step of decoding a speech signal to produce a decoded acoustic description of a word.

This can be seen from the fact that neither Smith nor Bahl '426 is capable of producing an acoustic description without text-based acoustic descriptions of the words. In Smith, if the text-based acoustic descriptions are removed, there are no pronunciations to combine and thus no pronunciations to score during speech recognition. Similarly, if the text-based acoustic descriptions of the words in Bahl '426 are removed, there would be no phonetic baseforms to use during fast match 102 and detailed match 106. Since neither Smith nor Bahl '426 can produce an acoustic description of a word without a text-based acoustic description, it is clear that neither is able to produce a decoded acoustic description from a speech signal. As such, claim 5 is patentable over the combination of Smith and Bahl '426.

Claims 6-8

Claims 6-8 were rejected under 35 U.S.C. § 103(a) as being obvious from Smith in view of Bahl '426 and further in view of Bahl et al. (U.S. Patent 6,377,921, hereinafter Bahl '921).

Bahl '921 provides a system for identifying transcription errors in text used for training a speech recognition system. Bahl '921 trains a set of acoustic models for acoustic units such as words, syllables, and phones. After the training is complete, a speech signal is aligned with its corresponding transcript using the trained models and a score is determined for each acoustic unit in the transcript. Instances of acoustic units that receive a low score from these models are then flagged and examined by a human operator to determine if the transcription is in error. Bahl '921 does not use a language model and does not decode a speech signal to produce a decoded acoustic description of a word.

Claims 6-8 depend indirectly from claim 1. As a result, they include the limitation to decoding a speech signal to produce a decoded acoustic description of a word pronounced by a user. The combination of Smith, Bahl '426 and Bahl '921 does not show or suggest this limitation.

As discussed above, Smith and Bahl '426 fail to show this limitation. Similarly, Bahl '921 fails to show or suggest a step of decoding a speech signal to produce an acoustic description of a word. Under Bahl '921, a speech signal is never decoded. Instead, the speech signal is applied to a known transcription of the speech. Since the transcription is already provided, there is no need to decode the speech signal. As such, Bhal '921 and the combination of Bahl '921 with Smith and Bahl '426 does not show or suggest the invention of claims 6-8.

In addition, in claim 6, generating a score for a decoded acoustic description includes generating a language model score for a sequence of syllable-like units. None of Smith, Bahl '426 or Bahl '921 show or suggest generating a language model score for a sequence of syllable-like units.

In the Office Action, it was asserted that Bahl '921 showed the production of a language model score for a sequence of

syllable-like units at column 2, lines 30-61. However, the cited section does not mention language models. As such, it does not show or suggest generating a language model score for a sequence of syllable-like units.

Smith also fails to discuss any type of language model. Only Bahl '426 actually mentions a language model. However, this is a word-based language model, not a syllable-like unit language model. As such, none of the cited references show or suggest a language model based on syllable-like units.

Since none of the cited references show or suggest generating a language model score for a sequence of syllable-like units and since none of the cited references produce a decoded acoustic description, claim 6 and claims 7 and 8, which depend therefrom, are patentable over Smith, Bahl '426 and Bahl '921.

Claims 9-11

Claims 9-11 were rejected under 35 U.S.C. § 103(a) as being obvious from Smith in view of Bahl '426 and further in view of Contolini et al. (U.S. Patent 6,233,553, hereinafter Contolini).

Contolini provides a method of selecting one pronunciation from a set of text-based pronunciations. Under Contolini, a plurality of text-based pronunciations are formed from the spelling of a word using a transcription generator. The top N pronunciations are provided to a speech recognition system, which applies a speech signal to the transcriptions representing each pronunciation. The transcription that scores highest is selected for storage. Contolini does not decode a speech signal to produce a decoded acoustic description, nor does it show the production of an acoustic model score for a syllable-like unit by generating acoustic model scores for each of a sequence of phonemes that form the syllable-like unit.

Claims 9-11 depend from claim 1 and as such include the limitation to decoding a speech signal to generate a decoded

acoustic description. None of Smith, Bahl '426 and Contolini show or suggest such a step. In particular, Contolini does not decode a speech signal to generate a decoded acoustic description.

In Contolini, a speech signal is applied against previously identified transcriptions to identify a score for each transcription. As a result, the speech signal is not being decoded to generate a decoded acoustic description. Instead, text-based acoustic descriptions are simply being scored based on the speech signal.

Since none of the cited references show a step of decoding a speech signal to generate a decoded acoustic description, claims 9-11 are patentable over the cited art.

In addition, none of Smith, Bahl '426 or Contolini show or suggest generating an acoustic model score for a sequence of syllable-like units by generating acoustic model scores for each of a sequence of phonemes that form the sequence of syllable-like units as found in claim 9.

In the Office Action, it was asserted that claim 4, column 7, line 6 and column 6, line 56 of Contolini show this limitation. Applicants respectfully dispute this assertion.

Claim 4 simply states that the sound units of claim 1 are acoustic units. Neither claim 1 nor claim 4 make any mention of syllable-like units or of determining an acoustic score for a syllable-like unit by determining acoustic scores for a sequence of phonemes that form the syllable-like units. Column 6, line 56 describes classes of phonemes including consonant and syllabic. This section does not suggest generating an acoustic score for a syllable-like unit by determining acoustic scores for a sequence of phonemes. Column 7, line 6 discusses filtering unlikely sequences of phonemes. It does not show or suggest determining an acoustic score for a syllable-like unit by generating acoustic

scores for each of a sequence of phonemes that form the syllable-like unit.

Since none of Smith, Bahl '426, or Contolini, show or suggest determining an acoustic score for a syllable-like unit by determining acoustic scores for phonemes that form the syllable-like unit, the combination of these references does not show or suggest claim 9.

Claims 12-17

Claims 12-17 were rejected under 35 U.S.C. § 102(a) as being anticipated by Gupta et al. (U.S. Patent 6,243,680 B1, hereinafter Gupta).

Gupta provides a system for selecting a pronunciation of a word for entry into a dictionary. Under Gupta, the text of a new word is first converted into a string of phonemes using a set of text-to-phoneme rules 412. These phonemes are placed in a graph structure with each branch in the structure being represented by a different phoneme. For each phoneme branch, a set of parallel branches are constructed, one for each phoneme that is similar to the initial phoneme in the graph. Additional parallel branches are then added for each allophone of each phoneme in the graph where an allophone is a particular pronunciation of a phoneme. Gupta then applies a set of speech utterances to the graph to score each path through the graph. The path with the highest score is selected as the pronunciation of the word.

Independent claim 12 provides a computer-readable medium having instructions for selecting a phonetic description of a word to add to a lexicon. These steps include receiving the text of the word and a speech signal representing a person's pronunciation of the word. The text of the word is converted into a text-based phonetic description while the speech signal is used to generate a speech-based phonetic description of the word.

Either the text-based phonetic description or the speech-based phonetic description is then selected for entry in the lexicon based on the correspondence between each phonetic description and the speech signal.

Gupta does not show or suggest the invention of claim 12 because it does not include a step of generating a speech-based phonetic description from a speech signal.

In the Office Action, it was asserted that the combination of preprocessing unit 402 and graph scoring unit 404 in Gupta showed this step. Applicants respectfully dispute this assertion.

Neither preprocessing unit 402 nor graph scoring unit 404 are capable of generating a phonetic description of a word. Instead, preprocessing unit 402 generates feature vectors for the speech signal (see col. 13, lines 36-56) and graph scoring unit 404 scores phonetic descriptions that have been generated from the text. Specifically, "[t]he feature vectors for each utterance are used to score the allophonic graph generated on the basis of the orthographic representation of the new word." (Gupta, col. 13, lines 61-63). Thus, graph scoring unit 404 does not generate a speech-based phonetic description, but simply scores the text-based phonetic descriptions proposed by graph generator 400.

The fact that Gupta does not produce a speech-based phonetic description can be seen clearly by removing all of the phonetic descriptions that are based on the text. If this is done, allophone graph generator 400 produces an empty graph. This empty graph is provided to graph scorer 404, which is then unable to function since it does not have any phonetic sequences to apply the speech signal against. If Gupta produced a speech-based phonetic description, this would not be true since the speech-based phonetic description would still be present even if the text-based phonetic descriptions were removed.

Since Gupta does not produce a speech-based phonetic description, it does not anticipate claim 12 or claims 13-17, which depend therefrom.

Claim 18

Claim 18 was rejected under 35 U.S.C. § 103(a) as being obvious from Gupta in view of Contolini.

Claim 18 depends from claim 12 and thus includes the limitation to generating a speech-based phonetic description of a word from a representation of a speech signal. Neither Gupta nor Contolini show this limitation.

In particular, Contolini does not show or suggest producing a speech-based phonetic description from a speech signal. Instead, Contolini simply applies a speech signal to previously defined phonetic descriptions in order to score each phonetic description. As such, Contolini does not produce a speech-based phonetic description.

Since neither Gupta nor Contolini produce a speech-based phonetic description, the combination of these two references does not show or suggest the invention of claim 18.

Claims 19-21

Claims 19-21 were rejected under 35 U.S.C. § 103(a) as being obvious from Schulze (U.S. Patent No. 6,167,369) in view of Gupta.

Schulze describes a system for determining the language of a document. To do this, Schulze generates a set of trigram models for each language, where each trigram model provides the probability a character trigram in the language. An input text is then divided into trigrams. The trigrams for the input text are scored using the models for each language to generate a total score for each language. Schulze does not show or suggest syllable-like units or forming n-grams of syllable-like units.

Independent claim 19 provides a speech recognition system with a language model that is trained through a series of

steps that include breaking each word in a dictionary into syllable-like units and for each word, grouping the syllable-like units into n-grams. The total number of n-gram occurrences in the dictionary is counted and for each n-gram, the total number of occurrences of the particular n-gram is divided by the total number of n-gram occurrences in the dictionary to form a language model probability for the n-gram.

The combination of Schulze and Gupta does not show or suggest the invention of claim 19. In particular, neither reference shows or suggests grouping syllable-like units found in dictionary words into n-grams.

In the Office Action, it was asserted that Schulze shows grouping syllable-like units from dictionary words into n-grams at column 1, line 29. Applicants respectfully dispute this assertion.

The cited section of Schulze discusses dividing an input sentence into individual character trigrams. It does not mention syllable-like units or forming n-grams from syllable-like units. Furthermore, it would not be obvious to use syllable-like units with Schulze. One goal of the Schulze system is to be able to identify the language of short text segments. If larger units were used instead of individual characters, there would be fewer n-gram probabilities calculated for short text segments thereby making it more difficult to identify the language of the text.

Thus, neither Gupta nor Schulze shows or suggest grouping syllable-like units into n-grams. As such, claim 19 and claims 20 and 21, which depend therefrom are patentable over the combination of Gupta and Schulze.

Claims 20 and 21 are additionally patentable over Schulze and Gupta. In claim 20, the dictionary words are broken into syllable-like units by preferring syllable-like units that occur more frequently in the dictionary than other syllable-like

units. Neither Schulze nor Gupta show or suggest this additional limitation.

In the Office Action, it was asserted that Schulze showed preferring syllable-like units that occur more often at column 12, lines 35-37. However, the cited section does not discuss syllable-like units or providing a preference for certain speech units when dividing a dictionary word into speech units. Instead, the cited section states that trigrams with low frequency counts are discarded from a trigram array.

Trigrams found in a corpus cannot be given a preference during the search for the trigrams. The reason for this is that there is no latitude in how trigrams are identified in a word. Under Schulze, the trigrams are identified simply by selecting three characters in a row in a word. Just because one three character sequence is later removed from the array does not influence the identification of the trigrams in the words. All of the trigrams are identified regardless of which ones are later discarded from the array.

Since the rules for identifying trigrams do not allow a preference to be applied so that one trigram is preferred over another during trigram identification, the cited section of Schulze cannot show or suggest preferring syllable-like units that occur more often in a dictionary over other syllable-like units when dividing words into syllable-like units. As such, the combination of Gupta and Schulze does not show or suggest the invention of claims 20 and 21.

Conclusion

In light of the above remarks, claims 1-21 are patentable over the cited art. Reconsideration and allowance of the claims is respectfully requested.

The Director is authorized to charge any fee deficiency required by this paper or credit any overpayment to Deposit Account No. 23-1123.

Respectfully submitted,

WESTMAN, CHAMPLIN & KELLY, P.A.

By: _____

Theodore M. Magee, Reg. No. 39,758
Suite 1600 - International Centre
900 Second Avenue South
Minneapolis, Minnesota 55402-3319
Phone: (612) 334-3222 Fax: (612) 334-3312

tmm

MARKED-UP VERSION OF REPLACEMENT PARAGRAPHS

Please replace the paragraph beginning at page 9, line 14 with the following:

A user may enter commands and information into the computer 110 through input devices such as a keyboard 162, a microphone 163, and a pointing device 161, such as a mouse, trackball or touch pad. Other input devices (not shown) may include a joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 120 through a user input interface 160 that is coupled to the system bus, but may be connected by other interface and bus structures, such as a parallel port, game port or a universal serial bus (USB). A monitor 191 or other type of display device is also connected to the system bus 121 via an interface, such as a video interface 190. In addition to the monitor, computers may also include other peripheral output devices such as speakers 197 and printer 196, which may be connected through an output peripheral interface 195~~0~~.

MARKED-UP VERSION OF REPLACEMENT CLAIMS

18. The computer-readable medium of claim 12 wherein the steps further comprise:

receiving an instruction to generate an audible pronunciation of a phonetic description previously added to the speech recognition lexicon;

retrieving the added phonetic description from the speech recognition lexicon; and

causing an audible pronunciation to be generated based on the retrieved phonetic description.

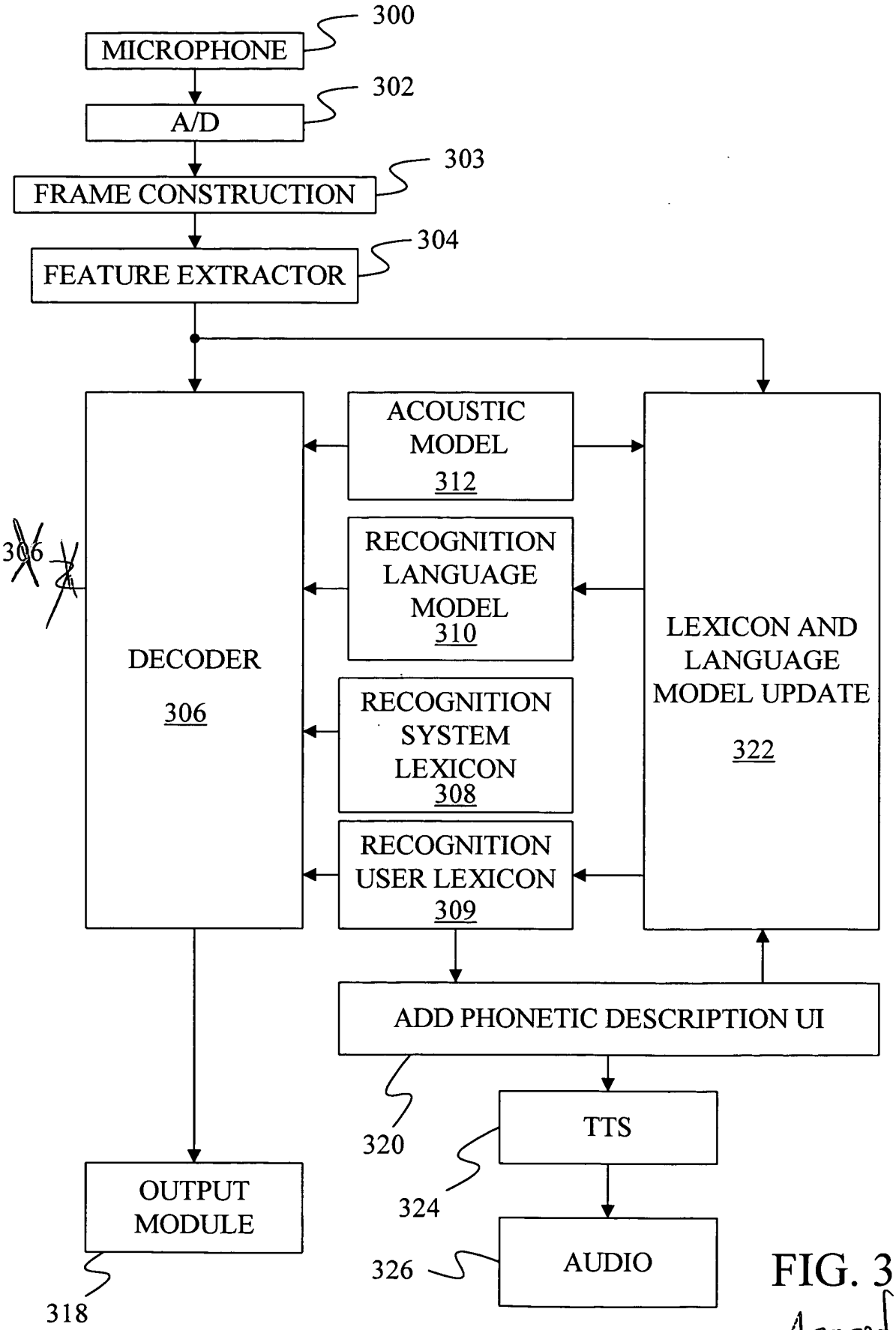

FIG. 3

Accepted Jul 12'03